Molecular Genetics and Genomics (2023) 298:549–554
https://doi.org/10.1007/s00438-023-01999-2

ORIGINAL ARTICLE

Effect of sample size on prognostic genes analysis in non-small cell lung cancer

Pingdong Li[1] · Haiyang Li[2] · Zhiyi Wan[3] · Yanan Lu[3]

Received: 10 November 2022 / Accepted: 15 February 2023 / Published online: 28 February 2023

Abstract

The identification of prognostic genes can help in the clinical management of non-small cell lung cancer (NSCLC). However, there is little overlap in the prognostic genes identified in different NSCLC studies. One reason for this may be the inadequate sample size. Here, the effect of sample size on prognostic genes analysis was investigated based on 515 stage II/III NSCLC cases from two cohorts detected by whole-exome sequencing. Prognostic genes analysis was repeatedly performed 100 times for each sample size level using random resampling methods. In stage II lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC) cases from the TCGA Pan-Lung Cancer cohort, the number of statistically significant prognostic genes first increased with sample size in a power law, then fluctuated steadily, and finally decreased slightly. The power law growth curves were also observed in stage III LUAD and LUSC cases from the TCGA Pan-Lung Cancer cohort and stage III Chinese LUAD cases from the OncoSG cohort. The correlation R^2 of the fitted power law growth curves were all greater than 0.99. In addition, at the sample size level where the number of prognostic genes peaked, the mean proportion of true prognostic genes in patients with stage II LUAD and LUSC was 28.32% and 23.12%, which could partly explain the little overlap in prognostic genes between reports. In conclusion, the number of prognostic genes takes a power law growth with the sample size in NSCLC, independent of histopathological subtype, race, and stage. These results also show how sample size affects the reliability of prognostic genes and will aid trial design for genomic mutation-based prognostic studies in NSCLC.

Keywords Non-small cell lung cancer · Sample size · Prognostic genes · Power law · Events number

Background

Lung cancer is the leading cause of cancer death in China and worldwide (Sung et al. 2021; Xia et al. 2022). Non-small cell lung cancer (NSCLC) accounts for the majority of lung cancer cases, with two major histological subtypes: lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC) (Nicholson et al. 2022). During the past decades, prognosis-related gene variants have been extensively identified in NSCLC (Wang et al. 2015; Campbell et al. 2016; Jiang et al. 2017; Meng et al. 2019; Caso et al. 2020; Chen et al. 2020). The identification of prognostic genes will contribute to a better understanding of the molecular features of NSCLC progression and guide clinical management. However, prognostic genes identified in different NSCLC studies rarely overlap (Jiang et al. 2017; Meng et al. 2019; Caso et al. 2020; Chen et al. 2020). It is often assumed that this phenomenon is due to certain differences in the cohorts used in different studies, such as age, gender, stage, and genetic background. However, the sample size is also important (Ein-Dor et al. 2005). In this study, we aimed to investigate the effect of sample size on prognostic genes analysis in NSCLC, which could contribute to the design of clinical trials for prognostic studies.

The number of genes in the human genome is about 20, 000, but sample sizes in different studies are usually in the tens to hundreds. Therefore, using all genes for prognostic analysis, we would be taking a high risk of overtraining.

Communicated by Shuhua Xu.

✉ Yanan Lu
nan0914@126.com

[1] Department of Otolaryngology, Head and Neck Surgery, Beijing Tongren Hospital, Capital Medical University, Beijing, China

[2] Department of Otolaryngology, People's Hospital of Beijing Daxing District, Beijing, China

[3] School of Biomedicine, Beijing City University, Beijing, China

One study showed that to achieve a typical overlap of 50% between two predictive gene lists, several thousand patients were needed in breast cancer studies (Ein-Dor et al. 2006). Some cancer genomic studies have shown that the small sample size may prevent the identification of robust prognosis-related genes (Brenton et al. 2005; Lønning et al. 2005). Therefore, the question is: how many samples are needed to generate a robust list of prognostic genes in NSCLC?

For prognostic prediction analyses, it should be ensured that the sample size is adequate in terms of the number of participants and outcome events relative to the number of predictor factors (Riley et al. 2019). Therefore, the duration of follow-up and the number of outcome events are also key factors influencing the prognostic analysis. Statistical power is determined by the number of events rather than the sample size itself. The longer the follow-up period, the smaller the sample size required to obtain the same number of events (Schober and Vetter 2018; In and Lee 2019).

Some formulas have been used to estimate the sample size required for survival analysis in clinical trials (Schoenfeld 1983; Hsieh and Lavori 2000). However, it is theoretically difficult to calculate the sample size required for prognostic genes analysis because the status of predictors (mutated genes) is unknown. Here, we investigated the effect of sample size on prognostic genes analysis using random resampling methods based on two real-world NSCLC cohorts.

Methods

Data

Clinical and genetic mutation data for NSCLC patients were obtained from the TCGA Pan-Lung Cancer cohort and OncoSG cohort (Campbell et al. 2016; Chen et al. 2020). After checking clinical information, a total of 515 NSCLC cases were included in the analysis, including 118 stage II LUAD cases, 82 stage III LUAD cases, 147 stage II LUSC cases, and 81 stage III LUSC cases from the TCGA Pan-Lung Cancer cohort, and 87 stage III LUAD cases from the OncoSG cohort.

Random sampling and prognostic analysis

To exclude confounding factors such as histopathological subtype and stage, we analyzed cases of the same stage in a single cohort separately. Prognostic genes associated with overall survival (OS) were identified by survival analysis stratified by the gene variant status using the "survGroup" function in maftools package (Mayakonda et al. 2018). *P* values in the survival analysis were determined using the log-rank test. Random resampling was conducted by randomly selecting n samples in the sample dataset. Random sampling and prognostic analysis were repeated 100 times for each sample size level. Mean values of prognostic gene counts were used for comparison and regression analysis.

Reliability analysis

First, the prognostic genes obtained in all cases were defined as true prognostic genes (TPGs). All prognostic genes obtained in each sampling analysis were then compared with the TPGs. The proportion of TPGs in each sampling was calculated as the number of overlapped TPGs / the number of all prognostic genes. The mean of the TPG proportions was used to assess the reliability of the prognostic genes obtained at different sample size levels. Reliability analyses were performed starting at the sample size level where the number of prognostic genes reached a plateau.

Statistical analysis

Statistical analysis was performed using R statistical software (V4.1.0). Tumor mutation burden (TMB) was calculated using maftools (Mayakonda et al. 2018) and analyzed by the Mann–Whitney *U* test. The threshold for statistical significance is 0.05. All results were presented as mean ± standard error.

Results

In stage II LUAD cases from the TCGA Pan-Lung Cancer cohort, the number of statistically significant prognostic genes increased with sample size in a power law with an exponent of 2.2256 until the sample size reached 70 (Fig. 1a). Then, the number of prognostic genes fluctuated steadily until the sample size reached 115 (Fig. 1a). In stage III LUAD cases from the TCGA Pan-Lung Cancer cohort, the power law growth curve was also observed (Fig. 1b). The number of prognostic genes still had not reached the plateau when the sample size reached 75 (Fig. 1b). In stage II LUSC cases from the TCGA Pan-Lung Cancer cohort, the power law growth curve with an exponent of 1.9654 was observed until the sample size reached 75 (Fig. 1c). Subsequently, the number of prognostic genes plateaued until the sample size reached 120 and then decreased slightly until the sample size reached 140 (Fig. 1c). In stage III LUSC cases from the TCGA Pan-Lung Cancer cohort, the number of prognostic genes increased with sample size in a power law with an exponent of 2.2627 until the sample size reached 65 (Fig. 1d). To achieve a 100% probability of obtaining statistically at least one statistically significant prognostic gene, the minimum sample sizes required were approximately 40 for stage II LUAD, 30 for stage III LUAD, 30 for stage II LUSC, and 30 for stage III LUSC, respectively.

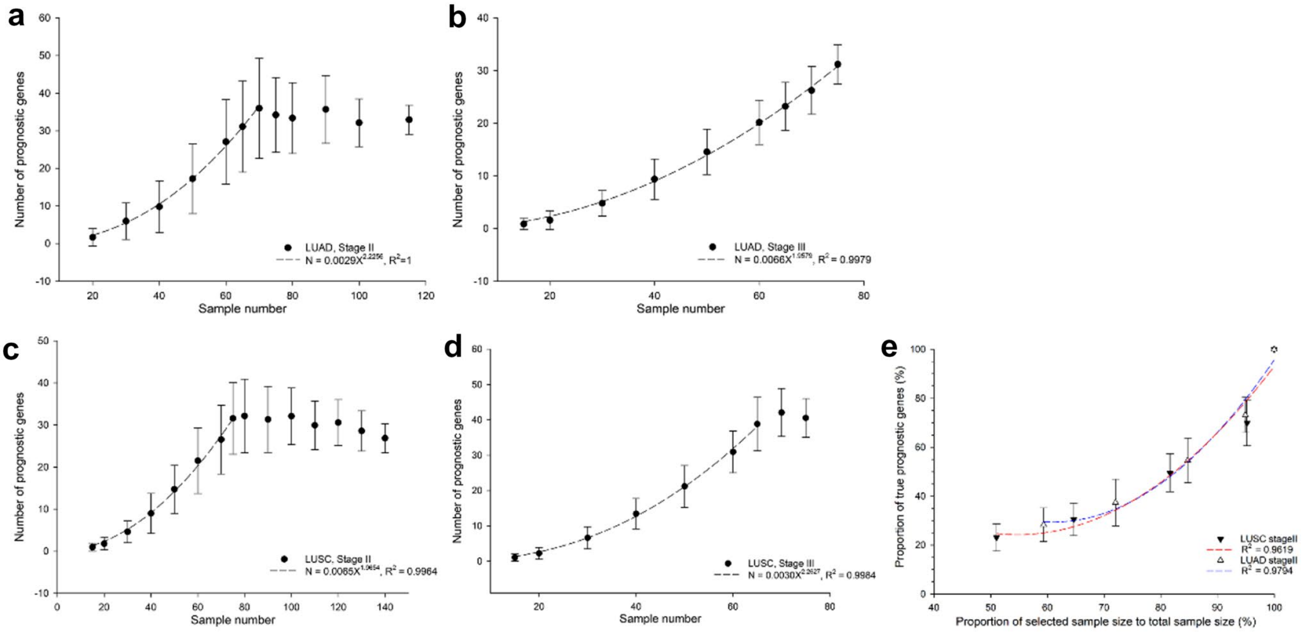

Fig. 1 Number of prognostic genes increases with sample size in LUAD (**a**, **b**) and LUSC (**c**, **d**) cases from the TCGA Pan-Lung Cancer cohort. The lines are the best-fit results for power law growth. **e** Proportion of true prognostic genes. The lines are the best-fit results for binomial function. Data are expressed as the mean ± standard error

The reliability of the prognostic genes obtained at different sample size levels was then evaluated. A total of 32 and 26 TPGs were identified in patients with stage II LUAD and LUSC, respectively (Table S1). At the sample size level where the number of prognostic genes reached the plateau, the mean of the TPG proportions was 28.32% in patients with stage II LUAD and 23.12% in patients with stage II LUSC (Fig. 1e). At the 95% total sample size level, the mean of the TPG proportions was 73.18% in patients with stage II LUAD and 69.96% in patients with stage II LUSC (Fig. 1e). In addition, the best-fit curves for the proportion of TPG relative to the proportion of sample size fit the binomial distribution and were very similar between stage II LUAD and LUSC (Fig. 1e).

The relationship between the number of prognostic genes and the events number was further analyzed. Using the events number corresponding to the samples number as the abscissa, similar curves were observed in both LUAD and LUSC (Fig. 2a). About 30–40 outcome events were required to reach the plateau of the number of prognostic genes (Fig. 2a). The growth curves of stage II LUAD and stage II/III LUSC nearly overlapped, while the curve of stage III LUAD was flatter. We speculated that this was due to the close TMB of stage II LUAD and stage II/III LUSC, while the TMB of stage III LUAD was lower. Comparative analysis of tumor mutation numbers confirmed this speculation (Fig. 2b).

We then analyzed the effect of genetic background on the number of prognostic genes. In stage III LUAD cases from the OncoSG cohort, all from Chinese LUAD patients, the number of prognostic genes also showed a power law increase with the sample size until the sample size reached 85 (Fig. 3a). However, the number of prognostic genes obtained under the same sample size was significantly lower than that of the Pan-Lung Cancer cohort (Fig. 3a). TMB analysis showed that the number of tumor mutations in stage III LUAD cases from the OncoSG cohort was significantly lower than that in stage III LUAD cases from the TCGA Pan-Lung Cancer cohort ($P < 0.001$, Fig. 3b).

Discussion

Prognosis is an important concern in the clinical management of NSCLC. Considerable effort has been devoted recently to OS prediction for NSCLC on the basis of genome sequencing (Wang et al. 2015; Campbell et al. 2016; Jiang et al. 2017; Meng et al. 2019; Caso et al. 2020; Chen et al. 2020). However, they are suffering from non-reproducibility among reports. Few studies have been conducted that specifically address the sample size requirements of prognostic studies. In here, we investigated the effect of sample size on prognostic genes using random resampling methods based on two cohorts, including LUAD cases from the OncoSG

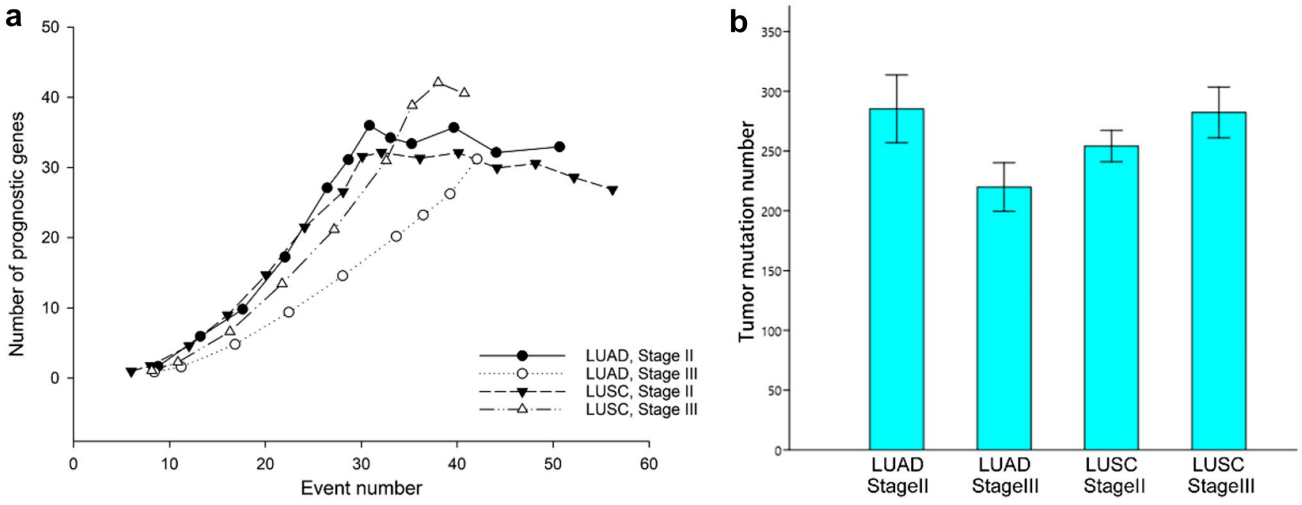

Fig. 2 Prognostic genes analysis in LUAD and LUSC. **a** Number of prognostic genes increases with events number. **b** Comparison of tumor mutational burden. Data are expressed as the mean ± standard error

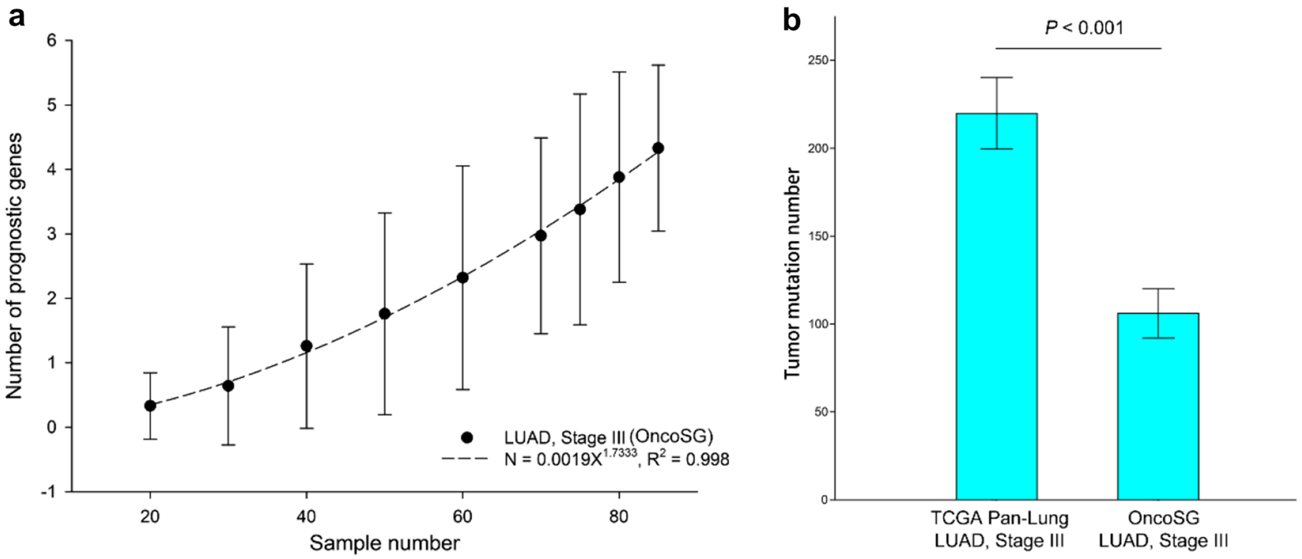

Fig. 3 Prognostic genes analysis in LUAD cases from the OncoSG cohort. **a** Number of prognostic genes increases with sample size. The line is the best-fit result for power law growth. **b** Comparison of tumor mutational burden. Statistical difference was analyzed by Wilcoxon test. Data are expressed as the mean ± standard error

cohort, and LUAD and LUSC cases from the TCGA Pan-Lung Cancer cohort. Patients in the TCGA Pan-Lung Cancer cohort were predominantly European Americans (Campbell et al. 2016), whereas patients in the OncoSG cohort were all Chinese (Chen et al. 2020).

In stage II NSCLC from the TCGA Pan-Lung Cancer cohort, including LUAD and LUSC, the number of prognostic genes first showed a power law increase with the sample size, then reached a plateau, and finally decreased slightly. Although the formula parameters of the power law curves were slightly different in different cohorts, the correlation R^2 of the fitted curves were all greater than 0.99. With the same sample size, the number of prognostic genes in Chinese LUAD was significantly lower than that in European Americans, possibly due to the lower TMB. However, the number of prognostic genes also increased with sample size in a power law. The prognostic analysis is based on time-to-event data (Moons et al. 2009). The number of outcome events is more critical than the sample size in prognostic analysis. Relative to the number of outcome events, the growth curves of the number of prognostic genes in the different cohorts were more similar. These results showed that the power law relationship between the number of prognostic genes and the sample size is common in NSCLC, independent of histopathological subtype, race, and stage.

Interestingly, our results also showed that the number of prognostic genes fluctuated steadily and decreased slightly as the sample size increased to a certain extent. Inclusion of more patients will detect more mutant genes and the number of prognostic genes will naturally increase. However, as the sample size increases, the statistical power becomes stronger and the mutant genes in NSCLC will also reach saturation. The balance of statistical power and the number of mutant genes may explain this phenomenon.

The number of prognostic genes that overlap between each sampling is difficult to assess directly due to the complexity of random sampling. Therefore, we used the proportion of TPGs to evaluate the reliability of the prognostic genes obtained at different sample size levels. Before the number of prognostic genes reached a plateau, the difference in the number and standard deviation of prognostic genes already implied that the reproducibility of prognostic genes is poor at this sample size level. Therefore, we only performed reliability analyses in stage II LUAD and LUSC cohorts with adequate sample sizes. The proportion of TPG relative to the sample size is consistent with a binomial distribution. At the sample size level where the number of prognostic genes peaked, the proportion of TPG in patients with stage II LUAD and LUSC averaged 28.32 and 23.12%, which could partly explain the little overlap in prognostic genes between reports. These results suggest that the effect of sample size on the reliability of prognostic genes is highly significant, even within the same cohort.

In summary, the number of prognostic genes follows a power law growth with sample size in NSCLC, independent of histopathological subtype, race, and stage. Our study suggests that at least 30–40 outcome events are required in NSCLC to reach the plateau of the number of prognostic genes. These results also show how sample size affects the reliability of prognostic genes and will contribute to the trial design of genomic mutation-based prognostic studies in NSCLC.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00438-023-01999-2.

Author contributions PL and YL designed the study and reviewed the manuscript. PL, HL, and ZW performed the data collection and statistical analysis. PL and ZW prepared the figures and drafted the manuscript. All the authors revised the manuscript and approved the final version.

Funding Not applicable.

Data availability All data generated or analyzed during this study are included in this published article.

Declarations

Conflict of interests The authors declare that they have no competing interests.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

Brenton JD, Carey LA, Ahmed AA, Caldas C (2005) Molecular classification and molecular forecasting of breast cancer: ready for clinical application? J Clin Oncol 23:7350–7360

Campbell JD, Alexandrov A, Kim J, Wala J, Berger AH, Pedamallu CS, Shukla SA, Guo G, Brooks AN, Murray BA, Imielinski M, Hu X, Ling S, Akbani R, Rosenberg M, Cibulskis C, Ramachandran A, Collisson EA, Kwiatkowski DJ, Lawrence MS, Weinstein JN, Verhaak RG, Wu CJ, Hammerman PS, Cherniack AD, Getz G, Artyomov MN, Schreiber R, Govindan R, Meyerson M (2016) Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas. Nat Genet 48:607–616

Caso R, Sanchez-Vega F, Tan KS, Mastrogiacomo B, Zhou J, Jones GD, Nguyen B, Schultz N, Connolly JG, Brandt WS, Bott MJ, Rocco G, Molena D, Isbell JM, Liu Y, Mayo MW, Adusumilli PS, Travis WD, Jones DR (2020) The underlying tumor genomics of predominant histologic subtypes in lung adenocarcinoma. J Thorac Oncol 15:1844–1856

Chen J, Yang H, Teo ASM, Amer LB, Sherbaf FG, Tan CQ, Alvarez JJS, Lu B, Lim JQ, Takano A, Nahar R, Lee YY, Phua CZJ, Chua KP, Suteja L, Chen PJ, Chang MM, Koh TPT, Ong BH, Anantham D, Hsu AAL, Gogna A, Too CW, Aung ZW, Lee YF, Wang L, Lim TKH, Wilm A, Choi PS, Ng PY, Toh CK, Lim WT, Ma S, Lim B, Liu J, Tam WL, Skanderup AJ, Yeong JPS, Tan EH, Creasy CL, Tan DSW, Hillmer AM, Zhai W (2020) Genomic landscape of lung adenocarcinoma in East Asians. Nat Genet 52:177–186

Ein-Dor L, Kela I, Getz G, Givol D, Domany E (2005) Outcome signature genes in breast cancer: is there a unique set? Bioinformatics 21:171–178

Ein-Dor L, Zuk O, Domany E (2006) Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. Proc Natl Acad Sci U S A 103:5923–5928

Hsieh FY, Lavori PW (2000) Sample-size calculations for the Cox proportional hazards regression model with nonbinary covariates. Control Clin Trials 21:552–560

In J, Lee DK (2019) Survival analysis: part II - applied clinical data analysis. Korean J Anesthesiol 72:441–457

Jiang Y, Huang Y, Du Y, Zhao Y, Ren J, Ma S, Wu C (2017) Identification of prognostic genes and pathways in lung adenocarcinoma using a bayesian approach. Cancer Inform 16:1176935116684825

Lønning PE, Sørlie T, Børresen-Dale AL (2005) Genomics in breast cancer-therapeutic implications. Nat Clin Pract Oncol 2:26–33

Mayakonda A, Lin DC, Assenov Y, Plass C, Koeffler HP (2018) Maftools: efficient and comprehensive analysis of somatic variants in cancer. Genome Res 28:1747–1756

Meng F, Zhang L, Ren Y, Ma Q (2019) The genomic alterations of lung adenocarcinoma and lung squamous cell carcinoma can explain the differences of their overall survival rates. J Cell Physiol 234:10918–10925

Moons KG, Royston P, Vergouwe Y, Grobbee DE, Altman DG (2009) Prognosis and prognostic research: what, why, and how? BMJ 338:b375

Nicholson AG, Tsao MS, Beasley MB, Borczuk AC, Brambilla E, Cooper WA, Dacic S, Jain D, Kerr KM, Lantuejoul S, Noguchi M, Papotti M, Rekhtman N, Scagliotti G, van Schil P, Sholl L, Yatabe Y, Yoshida A, Travis WD (2022) The 2021 WHO classification of lung tumors: impact of advances since 2015. J Thorac Oncol 17:362–387

Riley RD, Snell KI, Ensor J, Burke DL, Harrell FE Jr, Moons KG, Collins GS (2019) Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. Stat Med 38:1276–1296

Schober P, Vetter TR (2018) Survival Analysis and Interpretation of Time-to-Event Data: The Tortoise and the Hare. Anesth Analg 127:792–798

Schoenfeld DA (1983) Sample-size formula for the proportional-hazards regression model. Biometrics 39:499–503

Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F (2021) Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin 71:209–249

Wang R, Zhang Y, Pan Y, Li Y, Hu H, Cai D, Li H, Ye T, Luo X, Zhang Y, Li B, Shen L, Sun Y, Chen H (2015) Comprehensive investigation of oncogenic driver mutations in Chinese non-small cell lung cancer patients. Oncotarget 6:34300–34308

Xia C, Dong X, Li H, Cao M, Sun D, He S, Yang F, Yan X, Zhang S, Li N, Chen W (2022) Cancer statistics in China and United States, 2022: profiles, trends, and determinants. Chin Med J (engl) 135:584–590

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.